

# Approximation of the Scattering Amplitude using Nonsymmetric Saddle Point Matrices

Amber S. Robertson<sup>a</sup>, James V. Lambers<sup>a</sup>

<sup>a</sup>*Department of Mathematics, University of Southern Mississippi, 118 College Dr #5045,  
Hattiesburg, MS 39406, USA*

---

## Abstract

In this paper we examine iterative methods for solving the forward ( $A\mathbf{x} = \mathbf{b}$ ) and adjoint ( $A^T\mathbf{y} = \mathbf{g}$ ) systems of linear equations used to approximate the scattering amplitude, defined by  $\mathbf{g}^T\mathbf{x} = \mathbf{y}^T\mathbf{b}$ . Based on an idea first proposed by Gene Golub, we use a conjugate gradient-like iteration for a nonsymmetric saddle point matrix that is constructed so as to have a real positive spectrum. Numerical experiments show that this method is more consistent than known methods for computing the scattering amplitude such as GLSQR or QMR. We then demonstrate that when combined with known preconditioning techniques, the proposed method exhibits more rapid convergence than state-of-the-art iterative methods for nonsymmetric systems.

*Keywords:* nonsymmetric saddle point matrix, conjugate gradient method, scattering amplitude

---

## 1. INTRODUCTION

### 1.1. The Scattering Amplitude Problem

The core objective of this paper is to design and implement an iterative method for the solution of a system where the coefficient matrix is large, sparse, and nonsymmetric. The proposed method should be more efficient and robust than existing methods for solving such systems. One application in which such a system arises is in the computation of the *scattering amplitude*. The scattering amplitude, in quantum physics, is the amplitude of the outgoing spherical wave relative to that of the incoming plane wave [7]. It is useful when it is of interest to know what is reflected when a radar wave is impinging on a certain object. The scattering amplitude can be computed by

taking the inner product of the right hand side vector  $\mathbf{g}$  of the *adjoint system*

$$A^T \mathbf{y} = \mathbf{g} \quad (1)$$

and the solution  $\mathbf{x}$  of the *forward system*

$$A\mathbf{x} = \mathbf{b}. \quad (2)$$

Applications of the scattering amplitude come up in nuclear physics [1], quantum mechanics [14], and computational fluid dynamics (CFD) [4]. One particular application is in the design of stealth planes [1].

The scattering amplitude  $\mathbf{g}^T \mathbf{x} = \mathbf{y}^T \mathbf{b}$  creates a relationship between the right hand side of the adjoint system and the solution to the forward system in signal processing. The field  $\mathbf{x}$  is determined from the signal  $\mathbf{b}$  in the system  $A\mathbf{x} = \mathbf{b}$ . Then the signal is received on an antenna characterized by the vector  $\mathbf{g}$  which is the right hand side of the adjoint system  $A^T \mathbf{y} = \mathbf{g}$ , and it is expressed as  $\mathbf{g}^T \mathbf{x}$  [7]. We are interested in efficiently approximating the scattering amplitude. It is informative to look at methods that other researchers have used to solve this problem, which will be discussed below.

The solution of the linear system (2) is important for many applications beyond the scattering amplitude, such as in the numerical solution of PDE with non-self-adjoint spatial differential operators. This solution can be obtained in many different ways, depending on the properties of the matrix  $A$ . The  $LDL^T$  factorization can be used to solve some problems with a symmetric matrix or a *Cholesky factorization* can be used if the matrix is also known to be positive definite [9]. However, for large, sparse systems, an iterative method is preferred. The *conjugate gradient* method is the preferred iterative method for a symmetric positive definite matrix  $A$  [9]. However it is much more difficult to find this solution for a matrix that is not symmetric positive definite. In the case that we have a matrix that is not symmetric, we can use methods like the *biconjugate gradient (BiCG)* [3] and *generalized minimal residual (GMRES)* methods [17]. If we have a matrix that is symmetric but indefinite, *SymmLQ* [24, 19] is the iterative method of choice. Since the scattering amplitude depends on both the forward and adjoint problem, it makes sense to use methods that take both the forward and adjoint problems into account, like the *quasi-minimal residual (QMR)* [16] and *generalized least squares residual (GLSQR)* methods [25].

## 1.2. Approximation of the Scattering Amplitude

The method of this paper employs a conjugate gradient-like approach since, for large, sparse matrices, it is best to use an iterative approach, such as the conjugate gradient method [11] which is particularly effective for symmetric positive definite matrices. In particular, conjugate gradient has a very rapid convergence if  $A$  is near the identity either in the sense of a low rank perturbation or in the sense of the norm. In [9] it is stated that

**Theorem 1.** *If  $A = I + B$  is an  $n \times n$  symmetric positive definite matrix and  $\text{rank}(B) = r$  then the Hestenes-Stiefel conjugate gradient algorithm converges in at most  $r + 1$  steps.*

**Theorem 2.** *Suppose  $A \in \mathbb{R}^{n \times n}$  is symmetric positive definite and  $b \in \mathbb{R}$ . If the Hestenes-Stiefel algorithm produces iterates  $\mathbf{x}_k$  and  $\kappa = \kappa_2(A)$  then*

$$\|\mathbf{x} - \mathbf{x}_k\|_A \leq 2\|\mathbf{x} - \mathbf{x}_0\|_A \left( \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k,$$

where  $\|\mathbf{w}\|_A = \sqrt{\mathbf{w}^T A \mathbf{w}}$ .

It is also stated in [9] that the accuracy of  $\mathbf{x}_k$  is often better than this theorem predicts and that the conjugate gradient method converges very rapidly in the  $A$ -norm if  $\kappa_2(A) \approx 1$ , where  $\kappa_2(A)$  is the *condition number* of  $A$ , defined by

$$\kappa_2(A) = \|A\|_2 \|A^{-1}\|_2 = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$$

with  $\sigma_{\max}$  and  $\sigma_{\min}$  referring to the largest and smallest singular values, respectively.

Multiplying both sides of  $A\mathbf{x} = \mathbf{b}$  by  $A^T$  yields the normal equations with a symmetric matrix  $A^T A$  that is also positive definite when  $A$  is invertible. However, this approach is not conducive to solving the forward and adjoint problems simultaneously. Furthermore, a significant problem with using  $A^T A$  is that now the condition number in the two-norm is squared for  $A^T A$ . Since this increases the sensitivity of the matrix, possibly making it ill-conditioned, this paper explores an alternative approach. The idea is to transform the problems  $A\mathbf{x} = \mathbf{b}$  and  $A^T \mathbf{y} = \mathbf{g}$  into an equivalent system in which the matrix can be guaranteed to have real, positive eigenvalues, as well as eigenvectors that are in some sense orthogonal, which is then conducive to solution using

a conjugate gradient-like iteration. It is not necessarily symmetry that we seek, but we will have symmetry with respect to some inner product. To this end, we use an idea first proposed by Gene Golub in [5], and consider a nonsymmetric saddle point matrix that has the form

$$M = \begin{bmatrix} A^T W A & A^T \\ -A & 0 \end{bmatrix}.$$

As required by the definition of a nonsymmetric saddle point matrix, we assume that the matrix  $W$  is symmetric positive definite. The goal is to choose  $W$  so that we can guarantee  $M$  has real, positive eigenvalues. In this paper we will introduce the *nonsymmetric saddle point conjugate gradient* (NspCG) method to solve a nonsymmetric, large, sparse linear system, which will then allow us to compute the scattering amplitude. We will also use ILU preconditioning with NspCG, which gives rapid convergence compared to existing methods for solving such systems.

This paper is organized as follows. In Section 2 we discuss the known methods for solving a large linear system with iterative approaches to compute the scattering amplitude such as Bidiagonalization or least squares QR (LSQR), quasi minimum residual (QMR), and block generalized LSQR (GLSQR). In Section 3 we will introduce the method of this paper, NspCG. Section 4 will include an analysis of the numerical results. The preconditioning techniques and results can be found in Section 5. The conclusions and discussion of possible future work will be given in Section 6.

## 2. Methods for Solving the Linear Systems of the Forward and Adjoint Problems

### 2.1. QMR approach

The QMR approach [16, 7] is based on the spectral decomposition  $A = XDX^{-1}$ ; also the basis of the QMR approach is the unsymmetric Lanczos [9, 18] process which generates two sequences

$$V_k = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \dots & \mathbf{v}_k \end{bmatrix}$$

$$W_k = \begin{bmatrix} \mathbf{w}_1 & \mathbf{w}_2 & \dots & \mathbf{w}_k \end{bmatrix}$$

that are biorthogonal, meaning  $V_k^T W_k = I$ . We have the following relations:

$$AV_k = V_{k+1}T_{k+1,k}, \tag{3}$$

$$A^T W_k = W_{k+1}\hat{T}_{k+1,k}. \tag{4}$$

where the tridiagonal matrices

$$T_{k+1,k} = \begin{bmatrix} \alpha_1 & \gamma_1 & & & \\ \beta_1 & \alpha_2 & \gamma_2 & & \\ & \beta_2 & \ddots & \ddots & \\ & & \ddots & \ddots & \gamma_{k-1} \\ & & & \beta_{k-1} & \alpha_k \\ & & & & \beta_k \end{bmatrix} = \begin{bmatrix} T_{k,k} \\ \beta_k \mathbf{e}_k^T \end{bmatrix}$$

and

$$\hat{T}_{k+1,k} = \begin{bmatrix} \hat{\alpha}_1 & \hat{\gamma}_1 & & & \\ \hat{\beta}_1 & \hat{\alpha}_2 & \hat{\gamma}_2 & & \\ & \hat{\beta}_2 & \ddots & \ddots & \\ & & \ddots & \ddots & \hat{\gamma}_{k-1} \\ & & & \hat{\beta}_{k-1} & \hat{\alpha}_k \\ & & & & \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \hat{T}_{k,k} \\ \hat{\beta}_k \mathbf{e}_k^T \end{bmatrix}$$

have block structures in which  $T_{k,k}$  and  $\hat{T}_{k,k}$  are not necessarily symmetric.

The residual,  $\mathbf{r} = \mathbf{b} - A\mathbf{x}$ , in each iteration can be expressed as

$$\begin{aligned} \|\mathbf{r}_k\| &= \|\mathbf{b} - A\mathbf{x}_k\| \\ &= \|\mathbf{b} - A\mathbf{x}_0 - AV_k\mathbf{c}_k\| \\ &= \|\mathbf{r}_0 - V_{k+1}T_{k+1,k}\mathbf{c}_k\| \\ &= \|V_{k+1}(\|\mathbf{r}_0\|\mathbf{e}_1 - T_{k+1,k}\mathbf{c}_k)\| \end{aligned} \quad (5)$$

with a choice of  $\mathbf{v}_1 = \frac{\mathbf{r}_0}{\|\mathbf{r}_0\|}$  where  $\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$  and  $\mathbf{x}_k = \mathbf{x}_0 + V_k\mathbf{c}_k$ . We now have the quasi-residual  $\|\mathbf{r}_k^Q\| = \|\|\mathbf{r}_0\|\mathbf{e}_1 - T_{k+1,k}\mathbf{c}_k\|$ . Then we choose  $\mathbf{w}_1 = \frac{\mathbf{s}_0}{\|\mathbf{s}_0\|}$ , where  $\mathbf{s}_0 = \mathbf{g} - A^T\mathbf{y}_0$  and  $\mathbf{y}_k = \mathbf{y}_0 + \mathbf{w}_k\mathbf{d}_k$ . Then the adjoint residual is  $\|\mathbf{s}_k^Q\| = \|\|\mathbf{s}_0\|\mathbf{e}_1 - \hat{T}_{k+1,k}\mathbf{d}_k\|$ . The vectors  $\mathbf{c}_k$  and  $\mathbf{d}_k$  are the solutions of the least squares problems for minimizing  $\|\mathbf{r}_k^Q\|$  and  $\|\mathbf{s}_k^Q\|$ . So now the solutions can be defined as

$$\mathbf{x}_k = \mathbf{x}_0 + V_k\mathbf{c}_k \quad (6)$$

$$\mathbf{y}_k = \mathbf{y}_0 + U_k\mathbf{d}_k. \quad (7)$$

## 2.2. LSQR approach

In LSQR [7, 19], a truncated bidiagonalization is used in order to solve the forward and adjoint problems approximately. The bidiagonal factorization of

$A$  is given by  $A = UBV^T$  where  $U$  and  $V$  are orthogonal and  $B$  is bidiagonal. Thus the forward and adjoint systems can be written as

$$UBV^T \mathbf{x} = \mathbf{b} \quad (8)$$

$$VB^T U^T \mathbf{y} = \mathbf{g}. \quad (9)$$

Now we can solve (8) by solving the following two systems

$$B\mathbf{z} = U^T \mathbf{b} \quad (10)$$

$$\mathbf{x} = V^T \mathbf{z}, \quad (11)$$

and we can solve (9) by solving

$$B^T \mathbf{w} = V^T \mathbf{g} \quad (12)$$

$$\mathbf{y} = U^T \mathbf{w}. \quad (13)$$

We need to use the following recurrence relations in an iterative process to produce a bidiagonal matrix

$$AV_k = U_{k+1}B_k \quad (14)$$

$$A^T U_{k+1} = V_k B_k^T + \alpha_{k+1} \mathbf{v}_{k+1} \mathbf{e}_{k+1}^T \quad (15)$$

where  $V_k$  and  $U_k$  are matrices with orthonormal columns, and

$$B_k = \begin{bmatrix} \alpha_1 & & & & \\ \beta_2 & \alpha_2 & & & \\ & \beta_3 & \ddots & & \\ & & \ddots & \alpha_k & \\ & & & \beta_{k+1} & \end{bmatrix}.$$

Also we have that

$$\begin{aligned} A^T AV_k &= A^T U_{k+1} B_k = (V_k B_k^T + \alpha_{k+1} \mathbf{v}_{k+1} \mathbf{e}_{k+1}^T) B_k \\ &= V_k B_k^T B_k + \hat{\alpha}_k \mathbf{v}_{k+1} \mathbf{e}_{k+1}^T \end{aligned} \quad (16)$$

and

$$\hat{\alpha}_{k+1} = \alpha_{k+1} \beta_{k+1}.$$

Because  $B_k$  is bidiagonal, it follows that  $B_k^T B_k$  is symmetric and tridiagonal. It can be seen from (16) that (14) and (15) implicitly apply Lanczos

iteration to  $A^T A$ . Now this iterative process can be used to obtain the approximate solution to the forward and adjoint systems. We define the residuals at step  $k$  as

$$\mathbf{r}_k = \mathbf{b} - A\mathbf{x}_k \quad (17)$$

$$\mathbf{s}_k = \mathbf{g} - A^T \mathbf{y}_k \quad (18)$$

where

$$\mathbf{x}_k = \mathbf{x}_0 + V_k \mathbf{z}_k \quad \mathbf{y}_k = \mathbf{y}_0 + U_{k+1} \mathbf{w}_k.$$

The goal of the LSQR approach is to obtain an approximation that minimizes the norm of the residual. That is, the norm  $\|\mathbf{r}_k\| = \|\mathbf{b} - A\mathbf{x}_k\|$  is minimized. When working with the forward and adjoint problems, this approach is limited due to the relationship between the starting vectors

$$A^T \mathbf{u}_1 = \alpha_1 \mathbf{v}_1.$$

The above relationship does not allow  $\mathbf{v}_1$  to be chosen independently.

### 2.3. Generalized LSQR (GLSQR)

The GSLQR method [7, 25] overcomes the disadvantages of the LSQR method by choosing starting vectors  $\mathbf{u}_1 = \frac{\mathbf{r}_0}{\|\mathbf{r}_0\|}$  and  $\mathbf{v}_1 = \frac{\mathbf{s}_0}{\|\mathbf{s}_0\|}$  independently where, for an initial guess of  $\mathbf{x}_0$  and  $\mathbf{y}_0$ ,  $\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$  and  $\mathbf{s}_0 = \mathbf{g} - A^T \mathbf{y}_0$ . It is based on the factorizations

$$AV_k = U_{k+1} T_{k+1,k} = U_k T_{k,k} + \beta_{k+1} \mathbf{u}_{k+1} \mathbf{e}_k^T \quad (19)$$

$$A^T U_k = V_{k+1} S_{k+1,k} = V_k S_{k,k} + \eta_{k+1} \mathbf{v}_{k+1} \mathbf{e}_k^T \quad (20)$$

From the above we get that

$$\beta_{k+1} \mathbf{u}_{k+1} = A\mathbf{v}_k - \alpha_k \mathbf{u}_k - \gamma_{k-1} \mathbf{u}_{k-1} = \mathbf{c}_k \quad (21)$$

$$\eta_{k+1} \mathbf{v}_{k+1} = A^T \mathbf{u}_k - \delta_k \mathbf{v}_k - \theta_{k-1} \mathbf{v}_{k-1} = \mathbf{d}_k, \quad (22)$$

where the recursion coefficients  $\alpha_k$ ,  $\gamma_k$ ,  $\eta_k$ , and  $\theta_k$  are chosen to make  $U_k$  and  $V_k$  have orthonormal columns, which yields

$$\alpha_k = \mathbf{u}_k^T A \mathbf{v}_k, \quad (23)$$

$$\gamma_k = \mathbf{u}_{k-1}^T A \mathbf{v}_{k+1}, \quad (24)$$

$$\delta_k = \mathbf{v}_k^T A^T \mathbf{u}_k, \quad (25)$$

$$\theta_k = \mathbf{v}^T A^T \mathbf{u}_{k+1}. \quad (26)$$

We can define  $\mathbf{u}_{k+1} = \frac{\mathbf{c}_k}{\beta_k}$  and  $\mathbf{v}_k = \frac{\mathbf{d}_k}{\eta_k}$ , where  $\beta_k = \|\mathbf{c}_k\|$ , and  $\eta_k = \|\mathbf{d}_k\|$ . Now we have that

$$T_{k+1,k} = \begin{bmatrix} \alpha_1 & \gamma_1 & & & \\ \beta_2 & \alpha_2 & & & \\ & & \ddots & \ddots & \gamma_{k-1} \\ & & & \beta_k & \alpha_k \\ & & & & \beta_{k+1} \end{bmatrix} \quad S_{k+1,k} = \begin{bmatrix} \delta_1 & \theta_1 & & & \\ \eta_2 & \delta_2 & \ddots & & \\ & \ddots & \ddots & \theta_{k-1} & \\ & & \eta_k & \delta_k & \\ & & & & \eta_{k+1} \end{bmatrix}.$$

The residuals can be expressed as follows

$$\|\mathbf{r}_k\| = \|\mathbf{r}_0 - U_{k+1}T_{k+1,k}\mathbf{x}_k\| = \|\|\mathbf{r}_0\|\mathbf{e}_1 - T_{k+1,k}\mathbf{x}_k\|, \quad (27)$$

and

$$\|\mathbf{s}_k\| = \|\mathbf{s}_0 - V_k S_{k+1,k}^T \mathbf{y}_k - \alpha_{k+1} \mathbf{v}_{k+1} \mathbf{e}_{k+1}^T \mathbf{y}_k\|. \quad (28)$$

The solutions  $\mathbf{x}_k$  and  $\mathbf{y}_k$  are

$$\mathbf{x}_k = \mathbf{x}_0 + \|\mathbf{r}_0\| V_k T_{k,k}^{-1} \mathbf{e}_1 \quad (29)$$

$$\mathbf{y}_k = \mathbf{y}_0 + \|\mathbf{s}_0\| U_k S_{k,k}^{-1} \mathbf{e}_1. \quad (30)$$

### 3. Iterative Methods for Nonsymmetric Saddle Point Matrices

The matrix  $M$ , defined as follows

$$M \equiv \begin{bmatrix} A^T W A & A^T \\ -A & 0 \end{bmatrix}, \quad (31)$$

where  $A \in \mathbb{R}^{n \times n}$  is invertible and  $W$  is a symmetric positive definite matrix, is an example of a *nonsymmetric saddle point matrix*. It can be shown that  $\mathbf{x}^T M \mathbf{x} \geq 0$  for all  $\mathbf{x} \neq \mathbf{0}$ . To see this, we first let  $\mathbf{x} = \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix}$ . Then  $\mathbf{x}^T M \mathbf{x}$  can be written as

$$\begin{aligned} \mathbf{x}^T M \mathbf{x} &= \begin{bmatrix} \mathbf{y}^T & \mathbf{z}^T \end{bmatrix} \begin{bmatrix} A^T W A & A^T \\ -A & 0 \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix} \\ &= \mathbf{y}^T (A^T W A) \mathbf{y} - \mathbf{z}^T A \mathbf{y} + \mathbf{y}^T A^T \mathbf{z} \\ &= \mathbf{y}^T (A^T W A) \mathbf{y}. \end{aligned}$$

Now, if we let  $\mathbf{r} = A\mathbf{y}$  for any nonzero vector  $\mathbf{y}$ , then,  $\mathbf{r}^T = (A\mathbf{y})^T = \mathbf{y}^T A^T$ . since  $W$  is symmetric positive definite, we have that  $\mathbf{y}^T (A^T W A) \mathbf{y} = \mathbf{r}^T W \mathbf{r} > 0$ , since  $\mathbf{r}$  is nonzero due to  $A$  being invertible. On the other hand, if we assume  $\mathbf{y} = 0$ , then  $\mathbf{x}^T M \mathbf{x} = \mathbf{y}^T (A^T W A) \mathbf{y} = 0$ . That is, whether  $\mathbf{y}$  is nonzero or not,  $\mathbf{x}^T M \mathbf{x} = \mathbf{r}^T W \mathbf{r} \geq 0$ .



### 3.1. Ensuring a Real Positive Spectrum

We want to choose  $W$  so that the matrix  $M$  has a real positive spectrum, so it is suitable for a conjugate gradient-like iteration [15]. To make this choice we need to first define

$$\mathcal{M}(\gamma) \equiv \mathcal{J}p(M) = \mathcal{J}(M - \gamma I) = \begin{bmatrix} A^T W A - \gamma I & A^T \\ A & \gamma I \end{bmatrix},$$

where  $p$  is a polynomial of degree one in the form  $p(\zeta) = \zeta - \gamma$  for  $\gamma \in \mathbb{R}$  and

$$\mathcal{J} \equiv \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix}.$$

The goal here is to determine if there exists a symmetric positive definite matrix  $\mathcal{M}(\gamma)$  with respect to which  $M$  is symmetric, meaning that  $M$  is  $\mathcal{M}(\gamma)$ -symmetric if  $\mathcal{M}(\gamma)M = M^T \mathcal{M}(\gamma) = (\mathcal{M}(\gamma)M)^T$ .

Let us first define a generic nonsymmetric saddle point matrix

$$\mathcal{A} = \begin{bmatrix} \hat{A} & \hat{B}^T \\ -\hat{B} & \hat{C} \end{bmatrix}.$$

and then define  $\mathcal{M}(\gamma) = \mathcal{J}p(\mathcal{A})$ . We can use the following results from [15] to determine how to obtain a real positive spectrum:

**Lemma 3.** *Let the matrix*

$$\mathcal{J} \equiv \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix}$$

*be conformally partitioned with  $\mathcal{A}$ . Then*

- (1)  $\mathcal{A}$  is  $\mathcal{J}$ -symmetric, i.e.,  $\mathcal{J}\mathcal{A} = \mathcal{A}^T \mathcal{J} = (\mathcal{J}\mathcal{A})^T$ , and for any polynomial  $p$ ,
- (2)  $p(\mathcal{A})$  is  $\mathcal{J}$ -symmetric, i.e.,  $\mathcal{J}p(\mathcal{A}) = p(\mathcal{A}^T)\mathcal{J} = (\mathcal{J}p(\mathcal{A}))^T$ , and
- (3)  $\mathcal{A}$  is  $\mathcal{J}p(\mathcal{A})$ -symmetric, i.e.,  $(\mathcal{J}p(\mathcal{A}))\mathcal{A} = \mathcal{A}^T(p(\mathcal{A})^T)\mathcal{J} = (\mathcal{J}p(\mathcal{A})\mathcal{A})^T$ .

**Theorem 4.** *The symmetric matrix  $\mathcal{M}(\gamma)$  is positive definite if and only if*

$$\lambda_{\min}(\hat{A}) > \gamma > \lambda_{\max}(\hat{C}) \tag{32}$$

*where  $\lambda_{\min}$  and  $\lambda_{\max}$  denote the smallest and largest eigenvalues, respectively, and*

$$\|(\gamma I - \hat{C})^{-1/2} \hat{B}(\hat{A} - \gamma I)^{-1/2}\|_2 < 1. \tag{33}$$

A sufficient condition that makes  $\mathcal{M}(\gamma)$  positive definite can be derived from the above theorem.

**Corollary 5.** *The matrix  $\mathcal{M}(\gamma)$  is symmetric positive definite when (32) holds, and, in addition,*

$$\|\hat{B}\|_2^2 < (\lambda_{\min}(\hat{A}) - \gamma)(\gamma - \lambda_{\max}(\hat{C})). \quad (34)$$

For  $\gamma = \hat{\gamma} \equiv \frac{1}{2}(\lambda_{\min}(\hat{A}) + \lambda_{\max}(\hat{C}))$ , the right hand side of (34) is maximal and (34) reduces to

$$2\|\hat{B}\|_2 < (\lambda_{\min}(\hat{A}) - \lambda_{\max}(\hat{C})). \quad (35)$$

The preceding results lead to a simple approach to determining whether  $\mathcal{A}$  is suitable for a conjugate gradient-like iteration [15].

**Corollary 6.** *If there exists a  $\gamma \in \mathbb{R}$  so that  $\mathcal{M}(\gamma)$  is positive definite, then  $\mathcal{A}$  has a nonnegative real spectrum and a complete set of eigenvectors that are orthonormal with respect to the inner product defined by  $\mathcal{M}(\gamma)$ . In case  $\hat{B}$  has full rank, the spectrum of  $\mathcal{A}$  is real and positive.*

Using the previous results from [15], we obtain a simple criterion for determining whether the matrix  $M$  from (31) can be constructed in such a way as to satisfy the criterion in Corollary 6.

**Theorem 7.** *Let  $A$  be an invertible  $n \times n$  real matrix, and let  $W$  be a symmetric positive definite  $n \times n$  matrix that satisfies*

$$\sigma_{\min}(W) > \frac{2\kappa_2(A)}{\sigma_{\min}(A)}. \quad (36)$$

*Then the matrix  $M$  defined by*

$$M = \begin{bmatrix} A^T W A & A^T \\ -A & 0 \end{bmatrix}$$

*has real positive eigenvalues and eigenvectors that are orthogonal with respect to the inner product defined by  $\mathcal{M}(\gamma) = \mathcal{J}p(M)$ . That is, the above selection of  $W$  makes the matrix  $M$  suitable for a conjugate gradient-like iteration.*

Proof: We need to satisfy (32) with a proper selection of  $\gamma$ . Let

$$\gamma = \frac{1}{2}(\lambda_{\min}(A^T W A)),$$

based on Corollary 5. Because of how  $\gamma$  is defined,  $\gamma$  satisfies

$$\lambda_{\min}(A^T W A) > \gamma > 0, \quad (37)$$

which means (32) is also satisfied. Now we need to choose  $W$  so that (35) from Corollary 5 holds. We require

$$2\|A^T\|_2 < \lambda_{\min}(A^T W A), \quad (38)$$

or

$$2\sigma_{\max}(A) < \lambda_{\min}(A^T W A) \quad (39)$$

where  $\|A^T\|_2 = \|A\|_2$  is equal to the largest singular value of  $A$ ,  $\sigma_{\max}(A)$ . From the fact that  $A^T W A$  is symmetric positive definite, we obtain

$$\begin{aligned} \frac{1}{\lambda_{\min}(A^T W A)} &= \rho((A^T W A)^{-1}) \\ &= \|(A^T W A)^{-1}\|_2 \\ &\leq \|A^{-1}\|_2^2 \|W^{-1}\|_2 \\ &\leq \frac{1}{\sigma_{\min}(A)^2 \sigma_{\min}(W)}. \end{aligned}$$

Therefore, (39) is satisfied if

$$2\sigma_{\max}(A) < \sigma_{\min}(A)^2 \sigma_{\min}(W), \quad (40)$$

or, equivalently, if (36) is satisfied.  $\square$

It follows that the matrix  $W$  satisfies the requirements to make  $\mathcal{M}(\gamma)$  be symmetric positive definite and that  $\mathcal{A} = M$  has a real, positive spectrum from Corollary 6. This result makes the matrix suitable for a conjugate gradient-like iteration, as will be described below.

### 3.2. The Case $W = wI$

Let  $A = U\Sigma V^T$  be the SVD of  $A$ , where

$$U = [\mathbf{u}_1 \quad \cdots \quad \mathbf{u}_n], \quad V = [\mathbf{v}_1 \quad \cdots \quad \mathbf{v}_n]$$

and  $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_n)$ . In the case  $W = wI$  for some scalar  $w$ , the condition from Theorem 7 reduces to

$$w > \frac{2\kappa_2(A)}{\sigma_n}. \quad (41)$$

We now study the eigensystem of  $M$ . Let  $M\mathbf{x}_j = \lambda_j\mathbf{x}_j$  for  $j = 1, 2, \dots, 2n$ , where  $\mathbf{x}_j = [\mathbf{y}_j^T \ \mathbf{z}_j^T]^T$ . The form of  $M$  from (31), with  $W = wI$ , yields

$$wA^T A\mathbf{y}_j + A^T \mathbf{z}_j = \lambda_j \mathbf{y}_j, \quad (42)$$

$$-A\mathbf{y}_j = \lambda_j \mathbf{z}_j \quad (43)$$

for  $j = 1, 2, \dots, 2n$ . Substituting (43) into (42) yields

$$(1 - w\lambda_j)A^T \mathbf{z}_j = \lambda_j \mathbf{y}_j. \quad (44)$$

Multiplying through by  $A$  and applying (43), we obtain

$$(w\lambda_j - 1)AA^T \mathbf{z}_j = \lambda_j^2 \mathbf{z}_j.$$

It follows that each  $\mathbf{z}_j$  is a multiple of a left singular vector of  $M$ , and  $\lambda_j^2/(w\lambda_j - 1)$  is the square of the corresponding singular value. Furthermore, from (43), we find that  $\mathbf{y}_j$  is a multiple of a right singular vector of  $M$ .

We conclude that the eigenvectors  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{2n}$  of  $M$  are given by

$$\mathbf{x}_{2j-1} = \begin{bmatrix} -\lambda_j^+ \mathbf{v}_j \\ \sigma_j \mathbf{u}_j \end{bmatrix}, \quad \mathbf{x}_{2j} = \begin{bmatrix} -\lambda_j^- \mathbf{v}_j \\ \sigma_j \mathbf{u}_j \end{bmatrix}, \quad j = 1, 2, \dots, n, \quad (45)$$

with corresponding eigenvalues  $\lambda = \lambda_j^+, \lambda_j^-$  that satisfy the quadratic equation

$$\lambda^2 - \sigma_j^2 w \lambda + \sigma_j^2 = 0. \quad (46)$$

It can be shown directly from (45) and (46) that these eigenvalues are real and positive, and the corresponding eigenvectors linearly independent, if and only if  $w$  satisfies the weaker condition

$$w > \frac{2}{\sigma_n}, \quad (47)$$

which is consistent with the necessary and sufficient condition for  $\mathcal{M}(\gamma)$  to be positive definite given in Theorem 4.

### 3.3. Nonsymmetric Saddle Point Conjugate Gradient Method

Let  $A \in \mathbb{R}^{n \times n}$  be nonsymmetric. We will now introduce a Conjugate Gradient (CG) approach that solves the linear system  $A\mathbf{x} = \mathbf{c}$  by solving an equivalent system of the form  $M\mathbf{z} = \mathbf{b}$ , where

$$M \equiv \begin{bmatrix} A^T W A & A^T \\ -A & 0 \end{bmatrix}. \quad (48)$$

The matrix  $M$  is also not symmetric; however, the spectrum is entirely contained in the right half of the complex plane, due to the fact that  $\mathbf{x}^T M \mathbf{x} \geq 0$  for all  $\mathbf{x}$ . In the preceding discussion, we established that if  $W$  was chosen so as to satisfy the assumptions of Theorem 7, then  $M$  is diagonalizable with real, positive eigenvalues. Furthermore, the bilinear form  $(\mathbf{u}, \mathbf{v})_G = \mathbf{u}^T G \mathbf{v}$ , where  $G = \mathcal{M}(\gamma) = \mathcal{J}p(M)$ , is a proper inner product, as  $G$  is symmetric positive definite. It follows that  $M$  is  $G$ -symmetric and  $G$ -definite, meaning that  $(M\mathbf{u}, \mathbf{v})_G = (\mathbf{u}, M\mathbf{v}_G)$  for all  $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{2n}$ , and  $(\mathbf{u}, M\mathbf{u})_G > 0$  for all  $\mathbf{u} \neq \mathbf{0}$ .

Let the vectors  $\mathbf{p}$  and  $\mathbf{b}$  be defined by

$$\mathbf{b} = \begin{bmatrix} A^T W \mathbf{c} + \mathbf{d} \\ -\mathbf{c} \end{bmatrix}, \quad \mathbf{p} = \begin{bmatrix} \mathbf{d} \\ \mathbf{0} \end{bmatrix}, \quad (49)$$

where  $A\mathbf{x} = \mathbf{c}$ ,  $M\mathbf{z} = \mathbf{b}$ , and  $\mathbf{p}^T \mathbf{z} = \mathbf{d}^T \mathbf{x}$  is the scattering amplitude for given vectors  $\mathbf{c}$  and  $\mathbf{d}$  that represent the field and antenna, respectively. The following conjugate gradient method is based on a given inner product  $(\mathbf{u}, \mathbf{v})_G = \mathbf{v}^T G \mathbf{u}$  for solving the linear system  $M\mathbf{x} = \mathbf{b}$ .

#### Algorithm 3.1

**Input:** System matrix  $M$ , right hand side vector  $\mathbf{b}$ , inner product matrix  $W$ , initial guess  $\mathbf{x}_0$

**Require:**  $\mathbf{r}_0 = \mathbf{b} - M\mathbf{x}_0$

**for**  $i = 0, 1, \dots$  until convergence **do**

$$\alpha_i = \frac{(\mathbf{r}_i, \mathbf{p}_i)_G}{(\mathbf{p}_i, \mathbf{p}_i)_G}$$

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{p}_i$$

$$\mathbf{r}_{i+1} = \mathbf{r}_i - \alpha_i M \mathbf{p}_i$$

$$\beta_{i+1} = -\frac{(\mathbf{r}_{i+1}, \mathbf{p}_i)_G}{(\mathbf{p}_i, \mathbf{p}_i)_G}$$

$$\mathbf{p}_{i+1} = \mathbf{r}_{i+1} + \beta_{i+1} \mathbf{p}_i$$

**end for**

We have the inner product matrix  $G = \mathcal{M}(\gamma)M$  suggested by [15]. From [15], we see that this choice of  $G$  gives a working CG from the following lemma.

**Lemma 8.** *Suppose that the symmetric matrix  $\mathcal{M}(\gamma)$  is positive definite. Then Algorithm 3.1 is well defined for  $M$  and  $G = \mathcal{M}(\gamma)M$ , and (until convergence) the scalars  $\alpha_i$  and  $\beta_{i+1}$  can be computed as*

$$\alpha_i = \frac{(\mathbf{r}_i, \mathbf{r}_i)_{\mathcal{M}(\gamma)}}{(M\mathbf{p}_i, \mathbf{p}_i)_{\mathcal{M}(\gamma)}} \quad (50)$$

$$\beta_{i+1} = \frac{(\mathbf{r}_{i+1}, \mathbf{r}_{i+1})_{\mathcal{M}(\gamma)}}{(M\mathbf{r}_i, \mathbf{r}_i)_{\mathcal{M}(\gamma)}}. \quad (51)$$

With this choice of inner product matrix, it can be shown that the residuals computed using the preceding algorithm are, in some sense, orthogonal.

**Theorem 9.** *Each residual  $\mathbf{r}_k$  as defined in Algorithm 3.1 is orthogonal to all previous residuals with respect to  $\mathcal{M}(\gamma)$ , i.e.  $(\mathbf{r}_i^T, \mathbf{r}_j)_{\mathcal{M}(\gamma)} = 0$ , where  $i \neq j$ .*

Proof: We know that  $\mathbf{r}_{i+1} = \mathbf{r}_i - \alpha_i M\mathbf{p}_i$ . Let  $\alpha_i$  be defined as in (50). Also, we know that all of the search directions are orthogonal, i.e.  $\mathbf{p}_i^T \mathcal{M}(\gamma) M\mathbf{p}_j = 0$  for  $i \neq j$ . We want to show that  $\mathbf{r}_i \mathcal{M}(\gamma) \mathbf{r}_j = 0$ . This will be shown by induction, where the base case that we need to establish is

$$\mathbf{r}_{i+1}^T \mathcal{M}(\gamma) \mathbf{r}_i = 0, \quad i = 0, 1, \dots \quad (52)$$

To show this we use the definition of  $\alpha_i$  and the expression for the search directions in the above algorithm,  $\mathbf{r}_{i+1} = \mathbf{r}_i - \alpha_i M\mathbf{p}_i$ . Now we have that

$$\mathbf{r}_{i+1}^T \mathcal{M}(\gamma) \mathbf{r}_i = \mathbf{r}_i^T \mathcal{M}(\gamma) \mathbf{r}_i - \frac{\mathbf{r}_i^T \mathcal{M}(\gamma) \mathbf{r}_i}{\mathbf{p}_i^T M^T \mathcal{M}(\gamma) \mathbf{p}_i} \mathbf{p}_i^T M^T \mathcal{M}(\gamma) \mathbf{r}_i. \quad (53)$$

Reindexing the definition of the residual from the algorithm yields the following expression for  $\mathbf{r}_i$

$$\mathbf{r}_i = \mathbf{p}_i - \beta_i \mathbf{p}_{i-1}.$$

Substituting this into (53) gives

$$\mathbf{r}_{i+1}^T \mathcal{M}(\gamma) \mathbf{r}_i = \mathbf{r}_i^T \mathcal{M}(\gamma) \mathbf{r}_i - \frac{\mathbf{r}_i^T \mathcal{M}(\gamma) \mathbf{r}_i}{\mathbf{p}_i^T M^T \mathcal{M}(\gamma) \mathbf{p}_i} (\mathbf{p}_i^T M^T \mathcal{M}(\gamma) \mathbf{p}_i - \beta_i \mathbf{p}_{i-1}^T M^T \mathcal{M}(\gamma) \mathbf{p}_i) \quad (54)$$

Rearranging the last term in (54) yields

$$\mathbf{p}_{i-1}^T M^T \mathcal{M}(\gamma) \beta_i \mathbf{p}_i = \beta_i \mathbf{p}_i^T \mathcal{M}(\gamma) M \mathbf{p}_{i-1} = 0$$

because  $\mathcal{M}(\gamma)$  is symmetric, and we already know that the search directions  $\mathbf{p}_i$  are orthogonal with respect to  $\mathcal{M}(\gamma)$ . Now it is easy to see that the denominator in (54) and the last factor in the numerator cancel leaving

$$\mathbf{r}_{i+1}^T \mathcal{M}(\gamma) \mathbf{r}_i = \mathbf{r}_i^T \mathcal{M}(\gamma) \mathbf{r}_i - \mathbf{r}_i^T \mathcal{M}(\gamma) \mathbf{r}_i = 0.$$

Now we need to show that each residual is orthogonal to all previous residuals. We will do this by showing  $\mathbf{r}_i^T \mathcal{M}(\gamma) \mathbf{r}_{i-d} = 0$ , where  $d > 1$ . Our induction hypothesis is  $\mathbf{r}_{i-1}^T \mathcal{M}(\gamma) \mathbf{r}_{i-d} = 0$ . To show this, first shift the indices to get the expression

$$\mathbf{r}_i = \mathbf{r}_{i-1} - \alpha_{i-1} M \mathbf{p}_{i-1}.$$

Rearranging the recurrence relation for the search directions yields

$$\mathbf{r}_{i-d} = \mathbf{p}_{i-d} - \mathbf{p}_{i-1-d} \beta_{i-d}.$$

Using this expression for  $\mathbf{r}_i$  and  $\mathbf{r}_{i-d}$  we obtain

$$\begin{aligned} \mathbf{r}_i^T \mathcal{M}(\gamma) \mathbf{r}_{i-d} &= \mathbf{r}_{i-1}^T \mathcal{M}(\gamma) \mathbf{r}_{i-d} - \alpha_{i-1} \mathbf{p}_{i-1}^T M \mathcal{M}(\gamma) (\mathbf{p}_{i-d} - \mathbf{p}_{i-1-d} \beta_{i-d}) \\ &= \mathbf{r}_{i-1}^T \mathcal{M}(\gamma) \mathbf{r}_{i-d} - \alpha_{i-1} \mathbf{p}_{i-1}^T M^T \mathcal{M}(\gamma) \mathbf{p}_{i-d} + \\ &\quad \alpha_{i-1} \mathbf{p}_{i-1}^T M^T \mathcal{M}(\gamma) \mathbf{p}_{i-1-d} \beta_{i-d}, \end{aligned} \tag{55}$$

where

$$\mathbf{r}_{i-1}^T \mathcal{M}(\gamma) \mathbf{r}_{i-d} = 0$$

by the induction hypothesis. Now we are left with

$$\mathbf{r}_i^T \mathcal{M}(\gamma) \mathbf{r}_{i-d} = -\alpha_{i-1} \mathbf{p}_{i-1}^T M^T \mathcal{M}(\gamma) \mathbf{p}_{i-d} + \alpha_{i-1} \mathbf{p}_{i-1}^T M^T \mathcal{M}(\gamma) \mathbf{p}_{i-1-d} \beta_{i-d} = 0,$$

where both terms are 0 due to the orthogonality of the search directions.  $\square$

#### 4. Numerical Results

In this section, we will analyze the results from the methods described in this paper. These methods include QMR from Section 2.2, GLSQR from Section 2.3, and NspCG from Section 3.1. We have duplicated the results from [7] for GLSQR and QMR and will compare them against the results for our NspCG method.

We need to first define the following matrix,  $M$  is our nonsymmetric saddle point matrix

$$M = \begin{bmatrix} A^T W A & -A \\ A^T & 0 \end{bmatrix}$$

where  $W = wI$  is defined from (41). These examples are from [7].

#### 4.1. Example 1

This example uses the matrix created by `A=sprand(n,n,0.2)+speye(n)` in MATLAB where `n=100`. This creates a random sparse  $n \times n$  matrix, where 0.2 is the density of uniformly distributed nonzero entries, and adds this to the identity.

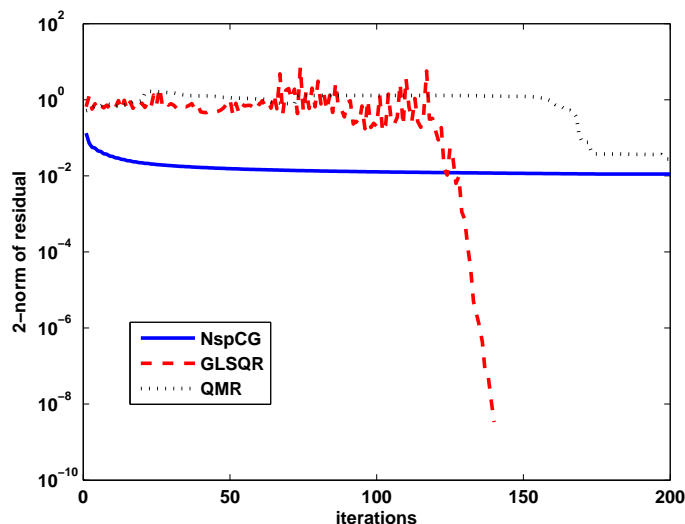


Figure 1: Example 1 with the matrix  $A$

In Figure 1 we see that at the beginning of the iteration NspCG reaches a better approximation in fewer iterations than either QMR or GLSQR. Although GLSQR eventually outperforms NspCG, it takes about 120 iterations before it shows any sign of convergence at all. Then it converges rapidly.

#### 4.2. Example 2

Example 2 uses the ORSIRR\_1 matrix from the Matrix Market collection, which represents a linear system used in oil reservoir modeling. This matrix can be obtained from <http://math.nist.gov/MatrixMarket/>.

We see that NspCG starts out with the lowest error in the 2-norm of the residual. Also we see that in both Figure 2 and Figure 1 that NspCG is more consistent than either GLSQR or QMR. Although QMR actually outperforms GLSQR and NspCG, it takes about 400 iterations to do so.



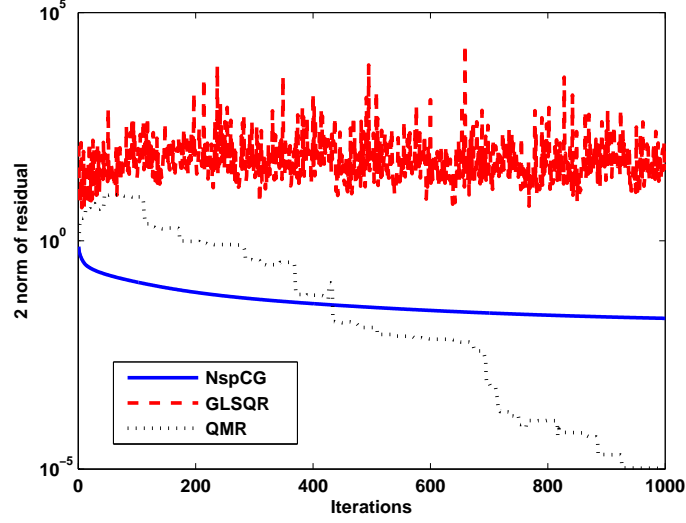


Figure 2: Example 2

#### 4.3. Example 3

First define the circulant matrix

$$J = \begin{bmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ 1 & & & 0 \end{bmatrix}.$$

Now the matrix used in this example  $A = 1e-3 * \text{sprand}(n, n, 0.2) + J$ , where  $n=100$ , can be constructed in MATLAB.

NspCG starts out steady and consistent again in this Figure 3 as we see in Figure 2 and Figure 1. Eventually, GLSQR converges, taking about 70 iterations to do so, while QMR fails to show any sign of convergence.

#### 4.4. Example 4

We need to first define

$$D_1 = \begin{bmatrix} 1000 & & \\ & \ddots & \\ & & 1000 \end{bmatrix} \in \mathbb{R}^{p,p} \quad D_2 = \begin{bmatrix} 1 & & \\ & 2 & \\ & & \ddots \\ & & & q \end{bmatrix} \in \mathbb{R}^{q,q}$$

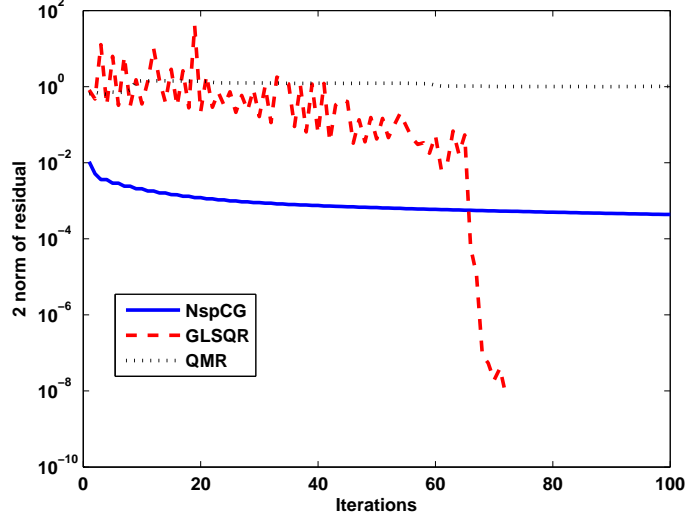


Figure 3: Example 3 with the matrix  $A$

where  $n = p+q$  and  $\Sigma = \text{diag}(D_1, D_2)$ . Now we can define  $A = U\Sigma V^T$ , where  $U$  and  $V$  are orthogonal matrices. For this example we use  $n = 100$  and  $D_1 \in \mathbb{R}^{90,90}$ . From 4 we see that NspCG starts off with the best approximation, but only for about 15 iterations. Then it is overtaken by GLSQR. Also, we can see that QMR fails to converge at all.

#### 4.5. Example 5

This example uses the same definition of  $D_1$ ,  $D_2$ , and  $A$  from Example 4. In this example we will let  $n = 100$  again, and  $D_1 \in \mathbb{R}^{50,50}$ . Figure 5 shows the same trend we have been seeing, that NspCG is more consistent at the beginning than any other method. At about 65 iterations GLSQR outperforms NspCG, and QMR fails to converge again.

#### 4.6. Example 6

This example uses the same definition of  $D_1$ ,  $D_2$ , and  $A$  from Example 4. In this example we will let  $n = 1000$  again, and  $D_1 \in \mathbb{R}^{600,600}$ . From Figure 6 we see that NspCG shows the best results for the first 600 iterations. GLSQR takes many iterations to converge in this case, and QMR does not converge at all.

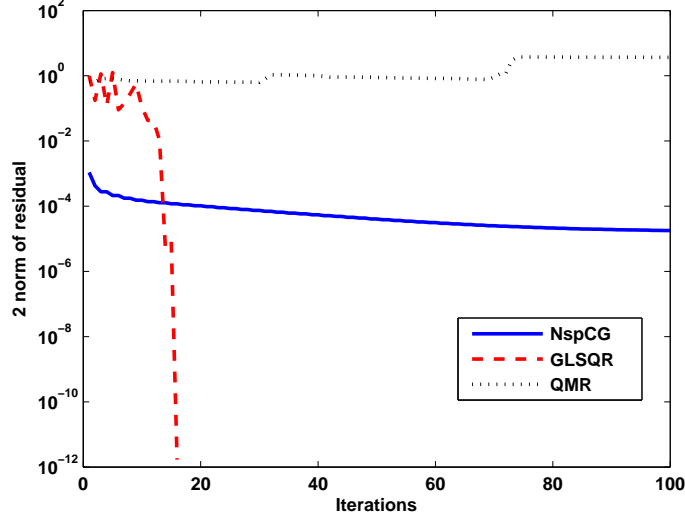


Figure 4: Example 4 with the matrix  $A$

## 5. Preconditioning

According to [9] conjugate gradient has very rapid convergence for a symmetric positive definite matrix  $A$  that is nearly identity. We need to apply *preconditioning* techniques to make our matrix  $M$  satisfy this criterion. The result will be that the original system is transformed into an equivalent system where the coefficient matrix is near identity. As we have seen previously with conjugate gradient, preconditioning techniques can be generalized to the nonsymmetric case. The goal is to apply *ILU* preconditioning [23], while taking into account the structure of the nonsymmetric saddle point matrix  $M$  defined in (31). The matrix  $W$  in the (1,1) block is assumed to be a symmetric positive definite matrix; therefore it has a Cholesky factorization  $W = GG^T$ . We can use the *QR* factorization

$$G^T A = QR$$

to obtain the factorization  $M = LU$ , where

$$L = \begin{bmatrix} R^T & 0 \\ -G^{-T}Q & G^{-T}Q \end{bmatrix}, \quad U = \begin{bmatrix} R & Q^T G^{-1} \\ 0 & Q^T G^{-1} \end{bmatrix}. \quad (56)$$

Let us define

$$C = G^T A \tilde{R}^{-1} \approx \tilde{Q},$$

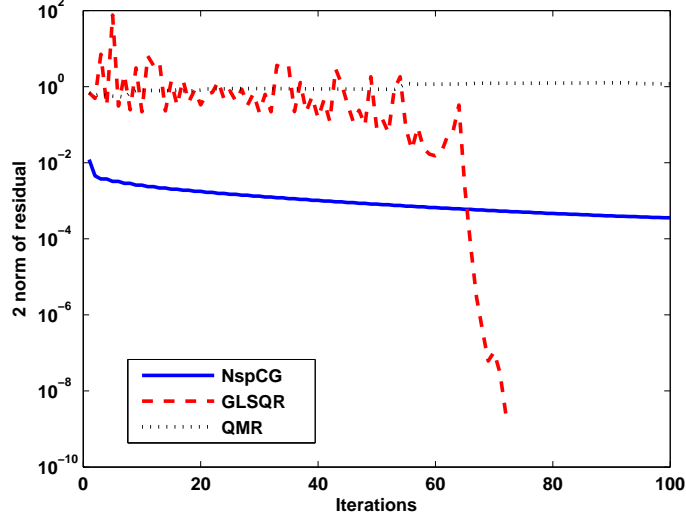


Figure 5: Example 5 with the matrix  $A$

where an incomplete  $QR$  factorization [20] is computed from the sparse matrix  $G^T A$  which gives  $G^T A \approx \tilde{Q}\tilde{R}$ . By finding

$$\tilde{L}^{-1} = \begin{bmatrix} \tilde{R}^{-T} & 0 \\ \tilde{R}^{-T} & \tilde{Q}G^T \end{bmatrix}, \quad \tilde{U}^{-1} = \begin{bmatrix} \tilde{R}^{-1} & -\tilde{R}^{-1} \\ 0 & G\tilde{Q}^{-T} \end{bmatrix} \quad (57)$$

it can be seen that the resulting preconditioned system matrix is given by

$$\tilde{L}^{-1}M\tilde{U}^{-1} = \begin{bmatrix} C^T C & -C^T C + C^T \tilde{Q} \\ C^T C - \tilde{Q}^T C & -C^T C + C^T \tilde{Q} + \tilde{Q}^T C \end{bmatrix}. \quad (58)$$

The above matrix has the structure similar to that of  $M$  from (31), therefore it is a nonsymmetric saddle point matrix that is near  $I$ .

### 5.1. Example 1

The following is Example 1 from the previous section with preconditioning.

### 5.2. Example 2

The following is Example 2 from the previous section with preconditioning. In Figure 8 NspCG converges very rapidly in only 10 iterations. QMR

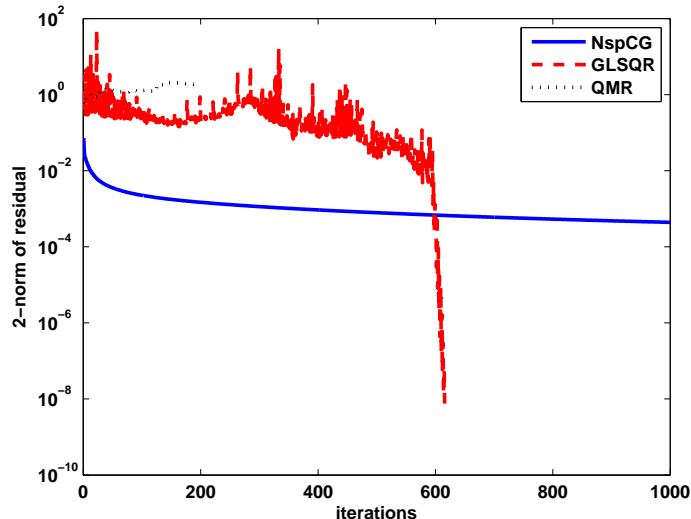


Figure 6: Example 6 with the matrix  $A$

takes over 200 iterations, but still doesn't reach the level of accuracy that NspCG achieves. GLSQR does not converge at all.

## 6. Conclusions and Future Work

The results from this paper show that the NspCG method is much more consistent and reliable than GLSQR or QMR. NspCG only takes a few iterations to make fairly significant progress while GLSQR takes many iterations in most cases, and QMR rarely makes any progress. If preconditioning is used with NspCG, as is usually done with a conjugate gradient method, we have provided evidence that it will dramatically accelerate convergence, compared to state-of-the-art iterative methods such as GMRES or BiCG that are typically used to solve such systems. These results support our hypothesis that more rapid convergence can be achieved by solving a system that, while still nonsymmetric, shares essential properties with symmetric positive definite matrices and therefore is more suitable for conjugate gradient-like iteration.

Future work will include relating the NspCG method to a quadrature rule, as in [6, 10], that can be used to compute the scattering amplitude without explicitly solving the forward or adjoint problem. This has been done in [7]

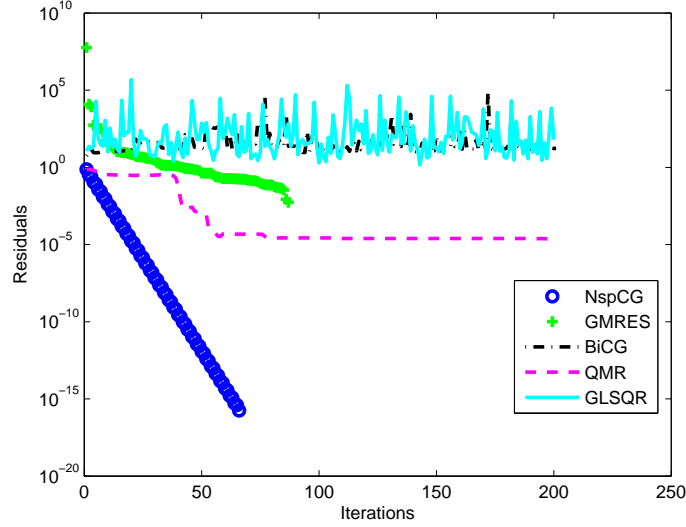


Figure 7: Example 1 with preconditioning

with the symmetric matrix

$$C = \begin{bmatrix} 0 & A^T \\ A & 0 \end{bmatrix}$$

in conjunction with block Lanczos iteration [8], but our goal is to achieve more rapid convergence. Furthermore, because the forward system  $A\mathbf{x} = \mathbf{b}$  is replaced with a system with twice as many unknowns and equations, it is essential to implement the iteration carefully so that the gain in convergence speed is not offset by the additional expense of each iteration. To that end, it is worthwhile to consider other choices for the matrix  $W$  instead of just a multiple of identity.

## References

- [1] Arnett, D. *Supernovae and Nucleosynthesis: An Investigation of the History of Matter, from the Big Bang to the Present*, Princeton University Press, (1996).
- [2] Björck, A. "A Bidiagonalization Algorithm for Solving Ill-posed System of Linear Equations". *BIT*, **41** (2001), pp. 659-670.

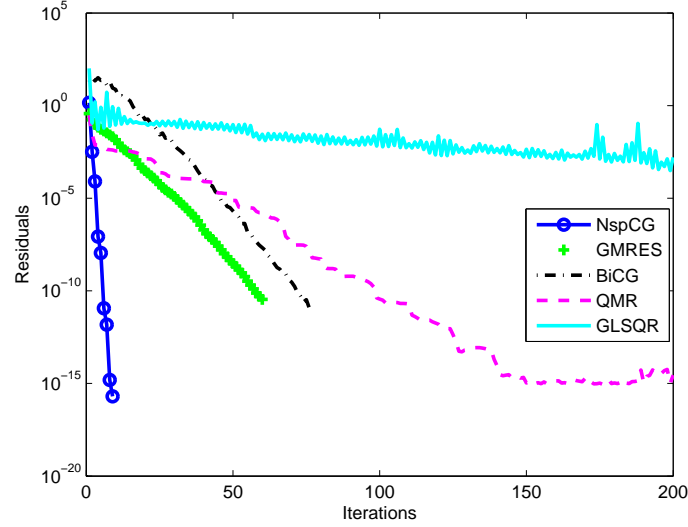


Figure 8: Example 2 with preconditioning

- [3] Brezinski, C., Redivo-Zaglia, M. "Look-Ahead in BiCGSTAB and Other Product-Type Methods for Linear Systems," *BIT*, **35** (1995), pp. 275-285.
- [4] Giles, M. B., Pierce, A. "An introduction to the adjoint approach to design", *Flow, Turbulence, and Combustion*, **65** (2000), pp. 393-415.
- [5] Golub, G. H., Lambers, J. V. Private communication, (November 6, 2007).
- [6] Golub, G. H., Meurant, G. "Matrices, Moments, and Quadrature". *Proceedings of the 15th Dundee Conference*, June-July (1993), *Longman Scientific and Technical*, (1994), pp. 105-156.
- [7] Golub, G. H., Stoll, M., Wathen, A. "Approximation of the Scattering Amplitude and Linear Systems". *ETNA*, **23** (2008), pp. 178-203.
- [8] Golub, G. H., Underwood, R. "The block Lanczos method for computing eigenvalues", *Mathematical Software III*, **7** (1977), pp. 361-377.
- [9] Golub, G. H., Van Loan, C.F.: *Matrix Computations*, The Johns Hopkins University Press (1996).

- [10] Golub, G. H., Welsch, J. "Calculation of Gauss Quadrature Rules" *Math. Comp.*, **23** (1969), pp. 221-230.
- [11] Hestenes, M., Stiefel, E. "Methods of Conjugate Gradients for Solving Linear Systems" *Journal of Research of the National Bureau of Standards* **49**(6) (1952).
- [12] Hnětynková, I., Strakoš, Z. "Lanczos Tridiagonalization and core problems". *Linear Algebra Appl.*, **421** (2007), pp. 243-251.
- [13] Lambers, J. V. "Matrices, Moments, and Quadrature".
- [14] Landau, L. D., Lifshitz, E. *Quantum Mechanics*, Pergamon Press, Oxford, (1965).
- [15] Liesen, J., Parlett, B. "On Nonsymmetric Saddle Point Matrices that allow Conjugate Gradient Iterations". *Numerische Mathematik*, **108** (2008), pp. 605-624.
- [16] Lu, J., Darmofal, L. "A quasi-minimal residual method for simultaneous primal-dual solutions, and superconvergent functional estimates", *SIAM J. Sci. Comput.*, **24** (2003), pp. 1693-1709.
- [17] Morgan, R.B. "A Restarted GMRES Method Augmented with Eigenvectors," *SIAM J. Matrix Anal. Applic.*, **16** (1995), pp. 1154-1171.
- [18] Morgan, R. B. "On Restarting the Arnoldi Method for Large Nonsymmetric Eigenvalue Problems", *Math Comp.*, **65** (1996), pp. 1213-1230.
- [19] Paige, C. C., Saunders, M. A. "Algorithm 583 LSQR: Sparse Linear Equations and Least Squares Problems", *ACM Trans. Math. Soft.*, **8** (1982b), pp. 195-209.
- [20] Papadopoulos, A.T., Duff, I. S., Wathen, A. J.: Incomplete Orthogonal Factorization Methods Using Givens Rotations II: Implementation and Results *BIT* **45**(1) (2005) 159-179.
- [21] Parlett, B. N., Nour-Omid, B. "The Use of a Refined Error Bound When Updating Eigenvalues of Tridiagonals", *Lin. Alg. and it's Applic.*, **34** (1980), pp. 31-48.



- [22] Parlett, B. N., Simon, H., Stringer, L. M., "On Estimating the Largest Eigenvalue with the Lanczos Algorithm", *Math. Comp.*, **38** (1982), pp. 153-166.
- [23] Saad, Y.: *Iterative methods for sparse linear systems*. PSW (1996).
- [24] Saunders, M. A. "Solution of Sparse Rectangular Systems," *BIT*, **35** (1995), pp. 588-604.
- [25] Saunders, M. A., Simon, H.D., Yip, E. L. "Two conjugate-gradient-type methods for unsymmetric linear equations", *SIAM J. Numer. Anal.*, **25** (1988), pp. 927-940.